Measuring Graded Membership: The Case of Color

Igor Douven
Sciences, Normes, Décision, Paris-Sorbonne University
igor.douven@paris-sorbonne.fr

Sylvia Wenmackers
Institute of Philosophy, University of Leuven
sylvia.wenmackers@hiw.kuleuven.be

Yasmina Jraissati
Department of Philosophy, American University Beirut
yasmine.jraissati@aub.edu.lb

Lieven Decock
Department of Philosophy, Vrije Universiteit Amsterdam
l.b.decock@vu.nl

Abstract

This paper considers Kamp and Partee's account of graded membership within a conceptual spaces framework and puts the account to the test in the domain of colors. Three experiments are reported which are meant to determine, on the one hand, the regions in color space where the typical instances of blue and green are located and, on the other hand, the degrees of blueness/greenness of various shades in the blue–green region as judged by human observers. From the locations of the typical blue and typical green regions in conjunction with Kamp and Partee's account follow degrees of blueness/greenness for the color shades we are interested in. These predicted degrees are compared with the judged degrees, as obtained in the experiments. The results of the comparison support the account of graded membership at issue.

Keywords: vagueness; graded membership; color; conceptual spaces; semantics

1 Introduction

Formal semantics is the study of truth and meaning with the help of logico-mathematical machinery. The main formal tool used in this area has traditionally been classical set theory. This has allowed researchers to model the truth of an atomic sentence $P(a)$ in terms of the item denoted by the constant a being included in the set denoted by the predicate P. Much research in formal semantics is concerned with generalizing this basic model to languages containing sentences of arbitrary complexity, and also containing various modal ("possibly," "necessarily"), temporal ("earlier," "later"), deontic ("must," "may"), and other operators.

Even the richest formal semantic models are, as a rule, idealized in that they assume predicates to either apply or not apply to any given item, thereby ignoring the fact that many natural-language predicates are vague to some extent and can thus apply—intuitively speaking—to intermediate degrees as well. An exception to this rule is the semantics presented in Kamp and Partee (1995), which explicitly aims at modeling vagueness of meaning in the context of prototype theory, specifically by defining a graded membership relation in terms of similarity to prototypical instances. This relation permits one to say things such as that the sentence "This apple is red" is true to a degree of .2 because the designated apple is a member to a degree of .2 of the set of red things.

Kamp and Partee acknowledged in their paper that the account of graded membership they propose is incomplete, given that in general it will fail to determine a unique graded membership function for all vague predicates in a language. It has recently been shown, however, that Kamp and Partee's proposal can be completed by embedding it within the conceptual spaces approach to concepts, which is enjoying growing popularity in cognitive science and cognitive psychology (see, e.g., Gärdenfors, 2000, 2014). Specifically, Decock and Douven (2014) show how combining Kamp and Partee's account with the conceptual spaces approach helps one to arrive at a unique graded membership relation, and Douven and Decock (2016) show how this relation can be used to define a unique relation of graded truth.

Since Tarski's (1935) seminal work on truth, two conditions have been widely accepted as adequacy criteria for any formal semantics, to wit, formal correctness and material adequacy. The former means that the semantics ought to be consistent—free of conflicting assignments of truth values—and the latter means that the semantics ought to capture the actual meanings of the terms in the language it aims to model, which in turn means that it ought to be in accordance with the linguistic behavior of competent speakers of the language. Douven and Decock (2016) proved the formal correctness of the conceptual spaces version of Kamp and Partee's semantics. But while it was recognized in Decock and Douven (2014) that, in this new version, the semantics has clear empirical content, the question of its material adequacy is still open.

The present paper reports empirical studies that aimed to take some first steps toward answering the question of material adequacy by applying the conceptual spaces version of Kamp and Partee's account to the domain of colors and querying whether, and if so to what extent, it makes the right predictions about people's use of (vague) color predicates. In particular, we obtained the data necessary to let Kamp and Partee's account (in the

said version) predict the degrees to which various color patches in the blue–green range are blue (or green), and then obtained further data on the degrees to which these patches were judged to be blue (or green) by human observers. The results will be seen to support Kamp and Partee's account. Needless to say, the empirical content of their account goes far beyond the color domain, let alone people's judgments of the degrees of blueness (greenness) of certain color patches. We comment on this and other limitations of the current experimental work in Section 3. The general discussion (Section 7) suggests some ways in which future work may overcome these limitations. In that section, we also contrast the conceptual spaces version of Kamp and Partee's semantics with the main extant psychological accounts of vague categorization.

2 Theoretical background

According to prototype theory, some instances of a concept are more representative for that concept than others, and the most representative instance is the prototype of the concept (see, e.g., Rosch, 1973, 1978, and Rosch & Mervis, 1975). When looking for a membership relation that admits of degrees, one might be tempted to equate degree of membership with similarity to prototype. For instance, one might think that the degree to which a given shade is red equals the similarity of that shade to prototypical red. However, this idea will not work, as Osherson and Smith (1981) pointed out long ago. For example, while shades of crimson are less typically red than shades of scarlet, they are nonetheless regarded as being red to the fullest degree (but see Barsalou, 1987).

For some, this observation has been a reason to dismiss as hopeless any attempt to define graded membership via similarity to prototypes. As Kamp and Partee argue, however, the observation at most militates against the thought that graded membership can be *equated* with similarity to prototype; it does not imply that similarity relations and prototypes cannot serve as a basis for an account of graded membership. In their 1995 paper, Kamp and Partee set out to show exactly how one can craft from those ingredients a relation of graded membership. They do so by considering a language with what they call "simple predicates" (predicates like "fish," "red," "tall," and others which have monolexemic expressions in English), and by constructing a semantics for this language that can account for the vagueness of those predicates (insofar as they are vague).

Kamp and Partee's semantics for their language consists of three parts: a domain of discourse, a partial model, and a set of completions. The domain of discourse is the set of all items the language allows one to talk about. The partial model assigns to each predicate in the language an extension and an anti-extension, where the union of these need not comprise the whole domain of discourse. The extension of a predicate consists of those items to which the predicate determinately applies, and the anti-extension consists of those items to which the predicate determinately fails to apply; but there may still be items remaining to which the predicate neither determinately applies nor determinately fails to apply. Completions are ways of handling these indeterminate cases. A completion assigns some of the indeterminate cases to the extension of the predicate and the remainder to the anti-extension.

Early attempts to deal with vagueness along these same lines considered *all* logically possible ways of splitting up the indeterminate cases of a predicate, declaring a sentence to be true if and only if it remains true no matter how the indeterminate cases are split up. The novelty of Kamp and Partee's proposal is that they only consider those ways of splitting up the indeterminate cases that respect orderings of similarity to prototype, meaning that if a is more similar to the P prototype than b is, then every completion that assigns b to the P extension should also assign a to the P extension, but not necessarily vice versa. To use their example, if both Alma and Bob are neither determinately adult nor determinately not adult, but Bob is *more* adult than Alma, then every admissible way of splitting up the indeterminate cases of adulthood that groups Alma with the clear cases should also group Bob with those cases, but there will be admissible ways of splitting up the indeterminate cases that group Bob with the clear cases but not Alma.

For the case of adulthood, there are only finitely many ways of splitting up the set of indeterminate cases (there are only finitely many people). A fortiori, there will be only finitely many completions. Then one can say that the degree of adulthood of a given person equals the proportion of completions that group that person with the clear instances. Trivially, every clear instance will on this definition receive a degree of adulthood of 1 and every clear non-instance will receive a degree of adulthood of 0. More interestingly, if Alma is grouped with the clear instances by, say, 60 percent of the completions, then her degree of adulthood equals .6, and if Bob is grouped with the clear instances by 80 percent of the completions, then his degree of adulthood equals .8.

This is the fundamental idea behind Kamp and Partee's definition of graded membership. The definition itself is slightly complicated by the fact that some predicates will have infinitely many associated completions. Therefore, the definition equates the degree of P membership of a not with the *proportion* of completions that group a with the clear P cases but rather with the (normed) *measure* of those completions (in the sense of measure theory; see Halmos, 1974). Kamp and Partee further argue for imposing a number of formal constraints on this measure. These constraints are meant to ensure that the measure is well-behaved in certain ways, such as that it assigns values in the $[0, 1]$ interval and that it is additive.

As stated in the introduction, however, Kamp and Partee acknowledge that this still does not yield a unique graded membership function. In their words,

> the constraints do not determine the function μ [i.e., the graded membership function] completely. Indeed, it is far from clear on what sorts of criteria a particular μ could or should be selected. (Kamp and Partee, 1995, p. 153)

Given the role that graded membership was supposed to play in their semantics, the noted shortcoming implies that the semantics still leaves wide open the truth values of many sentences in the language.

Decock and Douven (2014) argue that the elements of the above proposal all have natural interpretations in the context of a version of the conceptual spaces approach, and that, given those interpretations, the additional constraints to secure uniqueness of the graded membership function flow directly from the geometry of conceptual spaces. A conceptual space is a one- or multidimensional metric space whose dimension or dimensions represent fundamental qualities that items may have and whose associated metric

is used to measure similarities between items in the respect represented by the space. Concepts, in this view, are regions in a conceptual space. For example, color space can be thought of as a three-dimensional Euclidean space with one dimension representing hue, one dimension representing brightness, and one dimension representing saturation, and with the Euclidean metric defined on it measuring the color similarity between items. Color concepts are regions in this space. (For details on color space, see below.)

Before expanding on Decock and Douven's account, it is worth briefly elaborating on the provenance of conceptual spaces. The structure of a conceptual space is typically determined on the basis of similarity ratings. These ratings are given as input to one of several related statistical techniques that turn similarities, or more usually distances, into geometric objects. Most of these techniques go under the name of "multidimensional scaling." However, other techniques have been used as well; see, for instance, Castro, Ramanathan, and Chennubhotla (2013), who use nonnegative matrix factorization instead of multidimensional scaling to chart the structure of olfactory space. (See Clark, 1993, Appendix, and Gärdenfors, 2000, pp. 21–30, for more on this.)

The version of the conceptual spaces approach that Decock and Douven assume builds on an earlier version presented in Gärdenfors (2000, Chs. 3 and 4), which combines conceptual spaces with prototype theory and the technique of Voronoi tessellations. A Voronoi tessellation of a given space is a division of that space into non-overlapping cells such that each cell has a center and contains all and only those points in the space that lie no closer to the center of any other cell than to its own center. Points that lie equally close to two or more centers constitute the boundary lines. Formally, where S is an m-dimensional space, δ a metric defined on that space, and $\langle p_1, \ldots, p_n \rangle$ a sequence of pairwise distinct points in S, the region

$$v(p_i) \quad := \quad \{p \mid \delta(p, p_i) \leqslant \delta(p, p_j), \text{ for all } j \in \{1, \ldots, n\} \text{ with } j \neq i\}$$

is the *Voronoi polygon/polyhedron associated with* p_i. The elements of $\{v(p_i)\}_{1 \leqslant i \leqslant n}$ jointly constitute the *Voronoi diagram generated by* $\langle p_1, \ldots, p_n \rangle$ (see Okabe, Boots, Sugihara & Chiu, 2000, Ch. 2, for details). The left panel of Figure 2.1 shows a Voronoi tessellation on a two-dimensional Euclidean space, illustrating the above definition.

As Gärdenfors (2000, Ch. 3) argues, if the points representing prototypes in a given space are taken as the centers of a Voronoi tessellation, this tessellation carves up the space into cells which can be interpreted as concepts and, so interpreted, have various psychologically plausible properties. For instance, when a space is divided on this basis, the resulting cells are all convex (in the topological sense of the word; see Okabe et al., 2000, p. 58, for a proof). This is in line with empirical findings that suggest that concepts indeed correspond to convex regions (see, e.g., Sivik & Taft, 1994).

Decock and Douven take on board the amendment of Gärdenfors' version of the conceptual spaces approach that had been proposed in Douven, Decock, Dietz, and Égré (2013). In Gärdenfors' version, it is assumed that there is a unique prototype for each concept. However, at least if prototypes are understood in the sense of most representative instances, then in general that assumption can at best be an idealization. Surely there is not exactly one shade of red that best represents the concept of redness. It is therefore more realistic to say that instead of one prototypical *point* for red, we find in color space

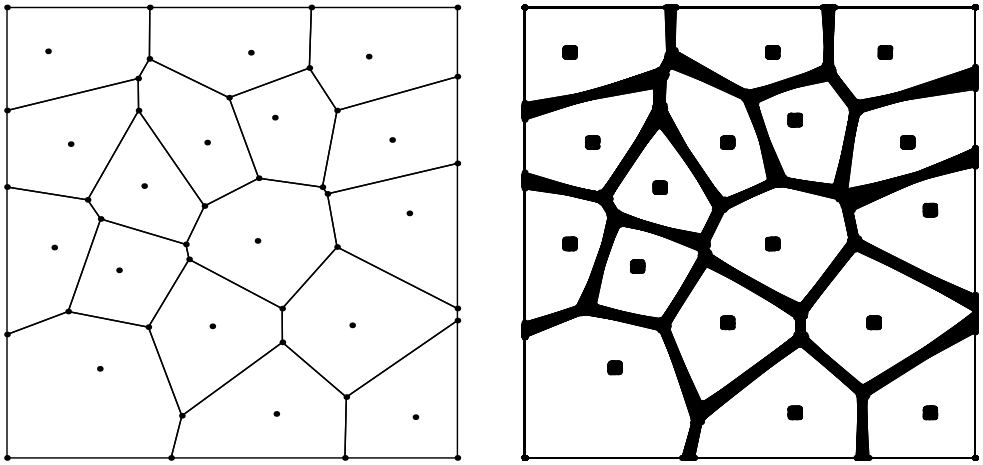

Figure 2.1: The left panel shows one of the "simple" Voronoi tessellations that make up the collated Voronoi tessellation shown in the right panel.

a prototypical *region* for red: a region of points that can all lay equal claim to being "maximally representative" for red (see in the same vein Regier, Kay, and Cook, 2005, p. 8390). But admitting as much requires a rethinking of the technique of Voronoi tessellations, given that, in standard definitions, such tessellations are generated by sets of points (the centers), not by sets of regions. To overcome this limitation, Douven et al. (2013) introduce (what they call) *collated Voronoi tessellations*, which are basically—as the name suggests—collations of ordinary Voronoi tessellations. More precisely, a collated Voronoi tessellation on a space overlays one over the other all ordinary tessellations on that space generated by a sequence of points which has as components one point from each prototypical region in the space.

Douven et al. (2013) show that the cells of a collated Voronoi tessellation still have the property of being convex, which is desirable if we interpret these cells as representing concepts (for the reason mentioned previously). Douven and colleagues also prove that, on the assumption that each prototypical region in a space is connected (again in the topological sense of the word), the boundary region of the corresponding collated Voronoi tessellation is "gapless." Psychologically speaking, the most important new feature of collated Voronoi tessellations is that their boundary regions have a certain "thickness," as the right panel of Figure 2.1 illustrates for a two-dimensional collated Voronoi diagram. To see why this is a desirable feature, note that we have no difficulty imagining a blue–green borderline case such that slightly altering its color along any of the three dimensions of color space will again result in a blue–green borderline case. So, speaking intuitively, a borderline case can be completely "surrounded" by other borderline cases. (Note that the preceding statement says "can," not "must": borderline cases on the edges are not completely surrounded by other borderline cases.) This fact is not adequately accounted for by Gärdenfors' version of the conceptual spaces framework, in which borderlines are all no more than one point "thick," whence "almost all" points neighboring a borderline

point (i.e., a point representing a borderline case) are not themselves borderline points but belong determinately to one or the other concept.

While it is an open issue how broadly applicable the conceptual spaces approach is, it has been quite generally accepted as a framework for modeling perceptual concepts, like color concepts (e.g., Helm, 1964; Shepard, 1964; Shepard & Carroll, 1966; Indow, 1988; Paramei, Izmailov, & Sokolov, 1991; Shepard & Cooper, 1992; Bosten, Robinson, Jordan & Mollon, 2005; Borg & Groenen, 2010, Ch. 4), auditory concepts (Petitot, 1989; Wang, Green, Samal & Yunusova, 2013), olfactory concepts (Castro, Ramanathan, & Chennubhotla, 2013), and shape concepts (Gärdenfors, 2000; Churchland, 2012). It has further been applied to action concepts (Gärdenfors, 2007; Gärdenfors & Warglien, 2012), event concepts (Gärdenfors & Warglien, 2012), moral concepts (Oddie, 2005), and scientific concepts like mass and acceleration (Gärdenfors & Zenker, 2011, 2013). Even if the conceptual spaces approach were limited to the aforementioned types of concepts, that would still make it sufficiently general to be of considerable interest.[1] The same would hold true for the conceptual spaces version of Kamp and Partee's account of graded membership summarized above. In the following, we set out to test this approach in the context of color concepts.

3 Overall design

In this paper, we offer an empirical test of the conceptual spaces version of Kamp and Partee's account of graded membership. In this account, the geometrical structure of a space—its dimensions and metric—in conjunction with the locations of the prototypical regions in that space determine uniquely, for any item and any concept representable in the space, the degree to which the item falls under the concept. As Decock and Douven (2014) show, if the geometrical structure of a space is known and if we also know where the prototypical regions in that space lie, we can effectively compute the degree to which a given item is a member of a given concept in the space. Thus, given a space for which such computations can be carried out, Kamp and Partee's account is testable if we can elicit people's judgments about the degrees to which items fall under various concepts in that space.

To make this more concrete, let (i) the geometry of color space be given, and suppose (ii) it were known where in this space are the prototypical regions for the colors corresponding to Berlin and Kay's (1969) basic color terms.[2] From this information, the present version of Kamp and Partee's account of graded membership permits one to predict, for any given shade, the degree to which people would judge it to belong to

[1] It is also sometimes said that the conceptual spaces approach, and geometric models of similarity generally, face the objection that similarity judgments can be asymmetric, as shown by Tversky (1977). See Gärdenfors (2000, p. 112 ff) for a response to this objection; see also Decock and Douven (2011). Independently, we note that, to the best of our knowledge, there is no evidence to support the claim that *color* similarity judgments can be asymmetric. So, even if geometric models of similarity were not tenable generally, there is no reason to doubt the geometric model of color similarity that will play an important role in the following (and that is widely used in color research).

[2] See Cook, Kay, and Regier (2005) for more recent data on basic color terms.

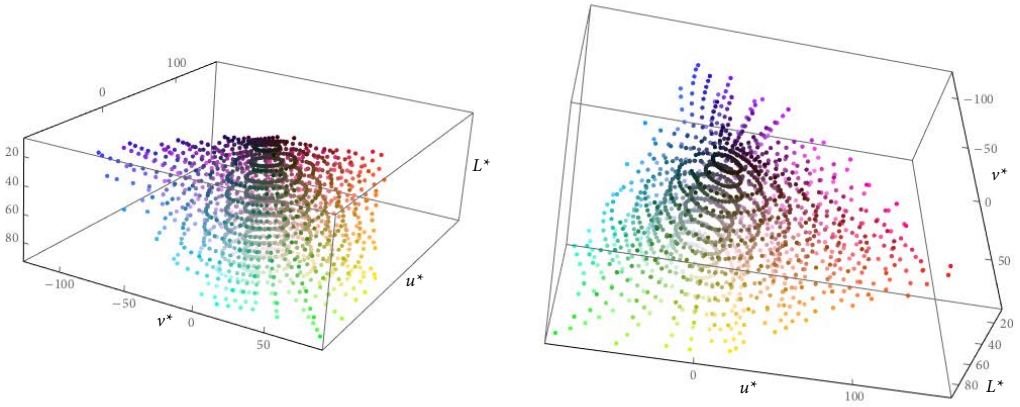

Figure 3.1: Different viewpoints of the set of 1625 chips from the Munsell laboratory placed in CIELUV space.

the various basic colors. It should be straightforward to compare such predictions with experimentally obtained color judgments, which would amount to a test of the designated account of graded membership.

As to (i), it is to be noted that color scientists regard different color spaces as appropriate for different viewing conditions. Specifically, the *Commission Internationale de l'Éclairage* has recommended the use of CIE 1976 $L^*u^*v^*$ space (or CIELUV space) for the characterization of colored displays on television or computer screens, and the use of the geometrically very similar CIE 1976 $L^*a^*b^*$ space (or CIELAB space) for the characterization of paints and colored surfaces (Malacara, 2002, pp. 86–90). The CIELUV and CIELAB spaces are designed to be perceptually uniform in the sense that pairs of color stimuli that human observers tend to perceive as equally different are mapped to pairs of points at equal distance in the space. While most color researchers consider them to be relatively close approximations of phenomenal color space, it is generally recognized that neither space is perfect in this respect (Fairchild, 2013, Ch. 10).

Because our experiments were conducted online, we will be working mostly in CIELUV space. For obvious reasons, online experiments are limited to stimuli that have RGB coordinates. Such stimuli are representable in CIELUV space but they can only cover part of it. Giving an impression of the RGB-representable part of CIELUV space, Figure 3.1 shows the locations in that part of a set of 1625 Munsell chips available from the website of the Munsell laboratory at the Rochester Institute of Technology (`http://www.rit.edu/cos/colorscience/rc_munsell_renotation.php`).

As to (ii), information on the locations in color space of prototypical color regions is limited. Berlin and Kay (1969), Sturges and Whitfield (1995), Benavente, Vanrell, and Baldrich (2006), and others report data on prototypicality judgments made by informants, but these data do not allow us to estimate the locations of prototypical color regions precisely enough for the purposes of our study.

The empirical work we undertook consisted of two main parts. The first—addressing the aforementioned point concerning (ii)—aimed to determine, or at least approximate, the locations in CIELUV space of the prototypical regions for two color categories, blue and green, and the second aimed to elicit degrees of blueness (or greenness) of various color shades in the blue–green region. From the results of the first part, Kamp and Partee's account of graded membership allowed us to calculate the degrees of blueness of the color shades used in the second part, which could then be compared with results from the second part in order to assess the predictive accuracy of Kamp and Partee's account.

Specifically, we conducted three experiments: Experiment 1 is a first step toward estimating the locations of the blue and green prototypical regions in CIELUV space. The first part of Experiment 2 takes the results of Experiment 1 as a starting point, with the aim of arriving at more accurate estimates of those prototypical regions. The second part of Experiment 2 elicits judgments to determine degrees of blueness (or greenness) of different color patches in three different series, ranging from blue to green. Experiment 3 controls for the possibility that the degrees of membership obtained in Experiment 2 might be biased by the specific protocol of the task, and uses different methods to elicit degrees of membership.

The three color series we used in Experiments 2 and 3 consisted of fourteen patches each. Although all three series go in a straight line (straight in CIELUV space, that is) from a determinately blue patch to a determinately green patch, the series were chosen so that two are relatively close to each other in CIELUV space—having the same direction, each at a constant brightness, with almost identical distances between patches in CIELUV space—and one has a different direction, ranges across different levels of brightness, and features a greater distance between the patches in CIELUV space than the first two series. This was done on purpose, to allow a precise assessment of the predictive accuracy of Kamp and Partee's account of graded membership. For the account to be accurate at least to a minimal degree, its predictions for the various degrees of blueness of the patches in the third series should at least match the judgments about those patches more closely than the judgments about the patches in the other series. But will Kamp and Partee's account (also) be able to discriminate between the two less well separated series? The predicted degrees for these series will inevitably be relatively close to each other, and it is reasonable to expect the same to hold for the observed degrees. Whether the predicted degrees for one of the two "close" series better match the observed degrees for *that* series than they match the observed degrees for the other thus constitutes a severe test for Kamp and Partee's account.

The methodology of our test of Kamp and Partee's account has some limitations, which are worth stating explicitly. First, Kamp and Partee offer what is meant to be a generally applicable semantics for languages with vague predicates. Our test is restricted to precisely two vague predicates, "blue" and "green." So, even if it passes our test, that still only gives *some* support to Kamp and Partee's semantics. But it may be support enough to warrant conducting further experimental studies, whether along the lines of the present one or otherwise. Moreover, the semantics may *fail* our test. That would show it to be materially inadequate, which would be an important (if perhaps undesirable) result.

Second, we conducted online experiments to test Kamp and Partee's semantics in the domain of color. Crowdsourcing services make it easy and relatively cheap for researchers to gather vast amounts of data. This explains the increasing popularity of this methodology in psychology and the social sciences. But while the methodology as such is not contested, one may have concerns over its use for perception studies. Traditionally, such studies have been conducted under standardized viewing conditions, mostly in laboratory environments. Internet-based experiments give the researcher no control over viewing conditions, and differences in those conditions due to participants' use of different monitors, operating systems, or web browsers are only to be expected. Such differences may seem especially troublesome for color experiments, where viewing conditions may matter even more than in experiments on, for instance, shape recognition. Still, in recent years a number of researchers have investigated the question of the validity of online color experiments by seeking to replicate in online studies the results of experiments that had been conducted in a controlled laboratory setting, and these replication attempts were generally successful; see in particular Moroney (2003), Sprow, Barańczuk, Stamm, and Zolliker (2009), Mylonas and MacDonald (2010), and Mylonas, Paramei, and MacDonald (2014). That being said, we *would* very much welcome a rerun of our experiments in the setting of a specialized laboratory.

Third, color similarity judgments are known to vary somewhat, both across people and across contexts of categorization. Hence, CIELUV space can at best capture the "average" person's perceptual color space in the "average" context of categorization.[3] Similarly, it is known that people can disagree about which shades are the most representative for a given color category. Ideally, then, one would want to proceed as follows: First predict, for each individual participant separately, and for each categorization context one is interested in, degrees of color membership of a range of color data on the basis of (i) a color space representing that participant's color similarity judgments in the given context, and (ii) the regions in that space containing the shades deemed prototypical for the various colors by the same participant given the same context. Next determine the degrees to which the color data belong to the various color categories, in the participant's judgment, and separately for the categorization contexts of interest. And finally, compare the thus obtained predicted and observed degrees of membership. While not practically impossible, the procedure outlined here would obviously require an enormous expenditure of time and effort, and probably also substantial financial resources. It is hardly prudent to follow this procedure in the absence of any empirical support for Kamp and Partee's semantics. This being so, it is best to start, less ambitiously, with somewhat simpler studies, such as the ones we are presenting, which in a way average not only over personal and contextualized perceptual spaces but also over judgments of typicality and judgments of graded membership.

[3] As mentioned, CIELUV space is taken to model color similarity judgments in contexts in which the stimuli are presented on a screen. There is evidence that differences in categorization task can lead to further systematic differences in similarity judgments (see, e.g., Nosofsky, 1987).

4 Experiment 1

To estimate the locations of the prototypical blue and green regions in CIELUV space, we proceeded in two stages. Experiment 1 constitutes the first stage, which was aimed at obtaining a first approximation of these locations.

4.1 Method

Participants

The total number of participants was 286. They had been recruited via the crowdsourcing interface CrowdFlower (`http://www.crowdflower.com`), which directed them to the Qualtrics platform (`http://www.qualtrics.com`) through which the survey was administered. In return for their cooperation, they were paid a small amount of money. All participants were from Australia, Canada, the United Kingdom, or the United States. Repeat participation was prevented.

We removed data from 2 participants who submitted incomplete surveys, 4 participants who indicated that they were colorblind, and 28 participants who failed a color sorting task that served as a quality control question, resulting in a sample of 252. The average age of these participants was 30 (SD = 11); 52 percent of them were female; 80 percent indicated university as their highest education level and 19 percent high school; the remaining participants had a lower education level. The participants included in the analysis spent on average 10 minutes on the survey (SD = 13 m).

The survey was in English (as were the surveys used in Experiments 2 and 3), but because this experiment only aimed at obtaining a *rough* estimate of the prototypical blue and green regions, we did not ask participants for their native language, nor did we exclude the fastest and slowest 5 percent responders (a common procedure in online experiments to enhance data quality). In the other experiments, where precision mattered more, nonnative speakers of English and fastest and slowest responders *were* excluded.

Materials and procedure

The materials of the present experiment consisted of $3 \times 27 = 81$ PNG images of $(2 \times 256)^2 = 512 \times 512$ pixels each (256^2 colors, 4 pixels per color). All images were cross-sections of RGB space with, for each of the R, G, and B axes, 27 equally spaced cross-sections normal to the axis. The images were produced by means of a homemade program, written in Object Pascal. By way of illustration, Figure 4.1 shows the middle (the fourteenth) image of each series.

In one part of the experiment, each participant was presented with a random selection of 25 of the 81 images, which were shown one after the other, each on a separate screen. Participants were asked to indicate, for each image they were shown, which spot in the image (if any) they considered to be typically blue; they were asked to indicate this by clicking on that spot in the image. The other part of the experiment was exactly the same, again using 25 randomly selected images, except that now the participants were asked to indicate which spot in each image (if any) they deemed typically green. Whether a participant received first the blue part and then the green part or the other way around was randomized per participant.

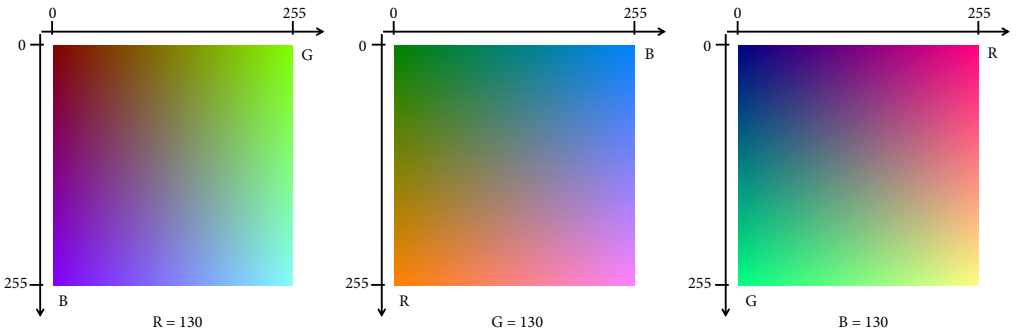

Figure 4.1: The middle images of the series of 27 cross-sections of RGB space along the R axis (left), the G axis (middle), and the B axis (right).

4.2 Results

In all, there were 4037 clicks indicating typically blue shades and 3741 clicks indicating typically green shades. The former are shown in CIELUV space in the upper left panel of Figure 4.2, the latter in the upper right panel. As is apparent from the figure, both for green and even more for blue, there were clicks on shades that probably anyone would agree are clearly *not* green or blue, respectively. For example, in response to the question concerning typical blue, some participants clicked on purple, red, or yellow shades. This will be partly noise, caused by participants clicking inadvertently (the Qualtrics software does not offer the opportunity to undo such errors) or by their not taking the task sufficiently seriously. On the other hand, it is also to be recognized that each image creates a context, in which for example even a purple shade may be judged to be typically blue by some participants. This was precisely the reason why we sliced up RGB space systematically along the three axes, so that we could average over all the clicks, thereby filtering out context effects as much as possible.

In averaging, we first calculated the medoids for both blue and green, where the medoid of a set of points is the point that minimizes the average distance to all points in the set (so, in the present case, the blue medoid is the "click" in the set of blue clicks that has the least average CIELUV distance to the points in that set). We then selected from the blue clicks the 50 percent that were closest (in CIELUV distances) to the blue medoid, and similarly from the green clicks the 50 percent that were closest to the green medoid. The lower left panel of Figure 4.2 shows the blue and green medoids, and the lower right panel shows the clicks for each color that remained after the selection.

5 Experiment 2

The experiment consisted of two parts: one that built on the results from Experiment 1 and that aimed to locate more precisely the green and blue prototypical regions in CIELUV space, and one that aimed to measure the degrees of blueness/greenness of three series of patches.

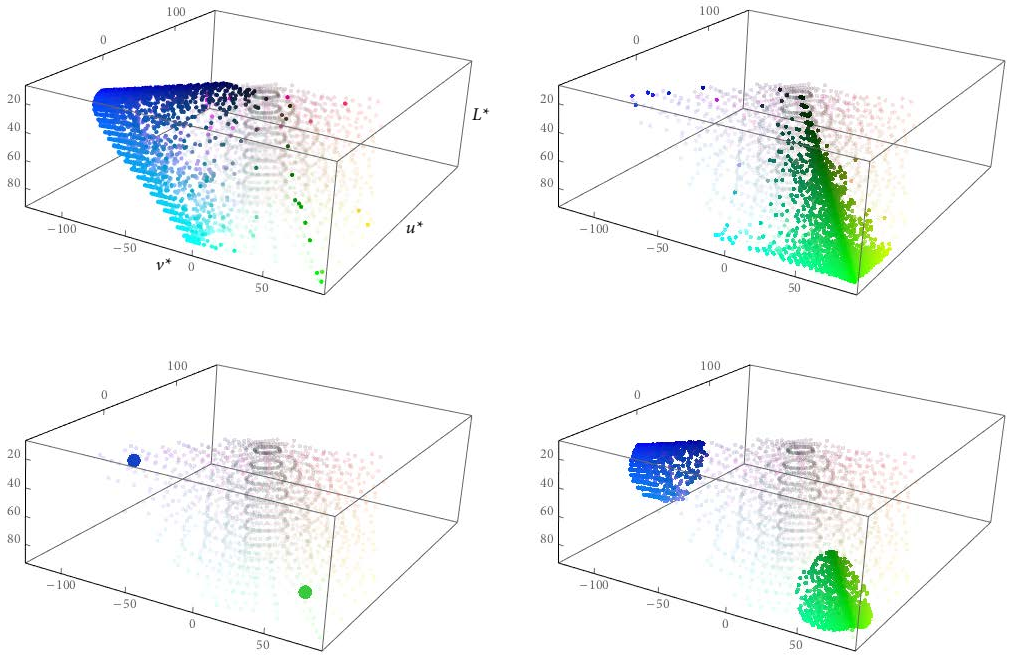

Figure 4.2: CIELUV space with responses for typical blue (top left); responses for typical green (top right); blue and green medoids (bottom left); 50 percent responses closest to medoids (bottom right).

5.1 Method

Participants

There were 394 participants in the study. They were again recruited via CrowdFlower, and the experiment was conducted on the Qualtrics platform. The participants received a small amount of money in compensation for their time. All participants were from Australia, Canada, the United Kingdom, or the United States. Repeat participation was prevented.

We removed data from 8 participants who indicated that they were colorblind and from 24 participants who indicated that they were nonnative speakers of English (the language of the survey), as well as the fastest and slowest 5 percent of responders.[4] We further removed the data from the 31 participants who failed the same color sorting task that had been used in Experiment 1 and from the one participant who indicated that he or she had not responded seriously. (Following a suggestion by Aust, Diedenhofen, Ullrich, ad Musch (2013), at the end of the survey the participants were asked whether

[4]We reran all analyses with fastest and slowest responders included and obtained qualitatively the same results as the ones reported in Section 5.2.

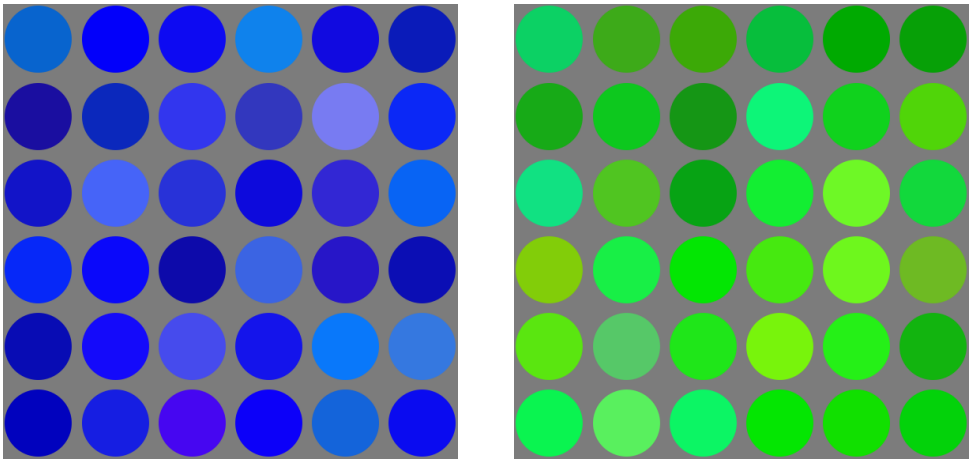

Figure 5.1: One of the four grids for blue (left) and one of the four grids for green (right).

they had responded seriously; it was emphasized that the answer to this question would not affect payment.) The final analysis was based on the responses of the remaining 290 participants.

These participants spent on average 390 seconds on the survey (SD = 120 s). Their mean age was 38 (SD = 12). Of these participants, 58 percent were female; 70 percent indicated university as their highest education level, 24 percent high school, and the remaining 6 percent a lower education level. Participants included in the analysis did not differ significantly from excluded participants in age, gender, or level of education.

Materials

The materials for the first part of the experiment—the part concerned with locating the prototypical regions—were 144 different points in CIELUV space randomly sampled from the 50 percent of clicks for blue that in Experiment 1 were nearest to the blue medoid, and 144 different points in CIELUV space randomly sampled from the 50 percent of clicks for green that in the same experiment were nearest to the green medoid. Each set of 144 points was divided into four subsets, and each subset was used to create a 6×6 grid of uniformly colored circles, where each circle had the color of one unique element of the subset. The circles were arranged against a uniformly gray background with RGB coordinates $\langle 124, 124, 124 \rangle$. Figure 5.1 shows one of the four grids for blue and one of the four grids for green. (The specific number of 144 was motivated strictly by practical concerns: it allowed us to create, for both blue and green, four grids of circular color patches such that the patches still appeared at a reasonable size on a computer screen.)

The materials for the second part of the experiment—the part concerned with eliciting degrees of blueness/greenness—consisted of three series of fourteen color patches each, all starting with a determinately blue patch and gradually transitioning to a determinately green patch. The series—simply referred to as "Series 1," "Series 2," and "Series 3" in the following—are shown in Figure 5.2. (For purposes of reporting the experimental results, throughout we assume the patches in each series to be numbered

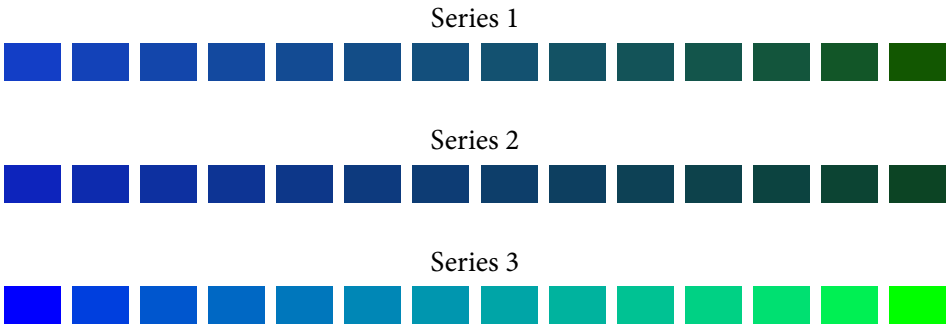

Figure 5.2: The three series that served as part of the materials.

1 through 14, from left to right as the series appear in Figure 5.2.) Series 1 goes from $\langle 31.9, -15.5, -92.1\rangle$ to $\langle 31.7, -27.9, 31.0\rangle$ in CIELUV coordinates (from $\langle 19, 62, 198\rangle$ to $\langle 18, 87, 1\rangle$ in RGB coordinates), with a CIELUV distance ($\Delta E^\star$) of 9.5 between each pair of adjacent patches. Series 2 goes from $\langle 25.0, -10.6, -87.9\rangle$ to $\langle 24.7, -20.8, 14.6\rangle$ (from $\langle 13, 36, 188\rangle$ to $\langle 12, 68, 36\rangle$ in RGB space), with a separation of $\Delta E^\star = 7.9$ between adjacent patches. Finally, Series 3 goes from $\langle 29.6, -11.5, -122.0\rangle$ to $\langle 87.8, -84.9, 87.2\rangle$, with a separation of $\Delta E^\star = 17.6$. In RGB space, Series 3 in effect goes from the "pure blue" point in that space, $\langle 0, 0, 255\rangle$ to the "pure green" point in that space, $\langle 0, 255, 0\rangle$.

Figure 5.3 depicts the three series in CIELUV space. Note that while each of the series goes in a straight line from the blue region to the green region, Series 3 is quite different from the other two, both in direction and in span. By contrast, Series 1 and 2 are rather close to each other, having the same direction and a comparable span.

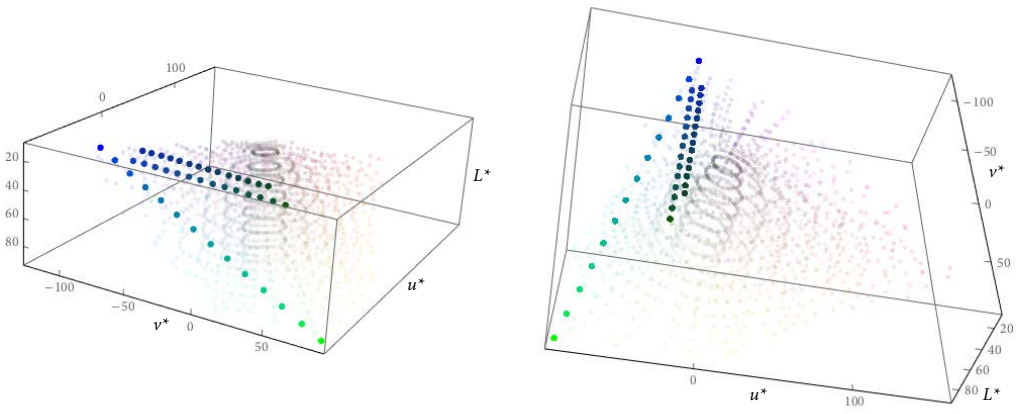

Figure 5.3: Different viewpoints of CIELUV space with Series 1–3 highlighted. The series with the greatest span is Series 3. Series 2 has the shortest span; it appears as the "uppermost" series in the left panel and as the "rightmost" series in the right panel.

Procedure

In the first part of the experiment, participants were shown the eight grids, one after the other, in an order randomized per participant, and participants were asked for each grid to click on the disk that struck them as being most typical for blue, respectively green. Participants were told that they could click on more than one disk, but that only their last click would be registered. They were also asked to take their time to consider alternative options before they continued to the next screen.

In the second part of the experiment, each participant was presented with two of Series 1–3, where the two series were chosen randomly per participant. Each chip of the chosen series was shown separately, in an individually randomized order, and the participants were asked to classify (by clicking one of two radio buttons) the color of each chip as either blue or green. It was explicitly noted at the beginning of this part of the experiment, and repeated on each screen, that the color shown might not fall clearly into either the green or the blue category, and participants were asked to choose the option that, in their opinion, *best* described the shown color.

5.2 Results

5.2.1 Determining predicted and observed degrees of membership

In the first part of the experiment, all participants clicked on each of the eight screens, yielding a total of 1160 judgments for typical blue and an equal number of judgments for typical green. This implies that the mean number of clicks received by the disks is $1160/144 \approx 8$ ($SD_{blue} = 11$; $SD_{green} = 17$). The maximum number of clicks received by a blue disk was 66; the maximum number of clicks received by a green disk was 116.

There is no obvious cutoff point in terms of number or percentage of clicks for counting a disk to exhibit a typical shade of blue, respectively green. Therefore, we chose to use various cutoff points, hoping that all choices would lead to qualitatively similar conclusions. Specifically, we decided to estimate the locations of the typical blue and typical green regions on the basis of (i) all 144 points in CIELUV space that had been sampled for each color in Experiment 1; (ii) the points with numbers of clicks greater than or equal to the first quartile; (iii) the points with numbers of clicks greater than or equal to the median number of clicks (second quartile); and (iv) the points with numbers of clicks greater than or equal to the third quartile.

To explain how we calculated degrees of membership for the colors in Series 1–3, we first recall that, on the conceptual version of Kamp and Partee's semantics, the degree to which an item belongs to a category is given by the measure of simple Voronoi tessellations that group that item with the category's prototypical region. In the present case, there are infinitely many such tessellations to be considered, given that the prototypical blue and prototypical green regions contain infinitely many points (whichever of the above choices we make). It is computationally impossible to consider all of these. However, we can *approximate* degrees of membership by randomly sampling from the simple Voronoi tessellations and determining degrees of membership on the basis of the resulting sample. This is how we proceeded.

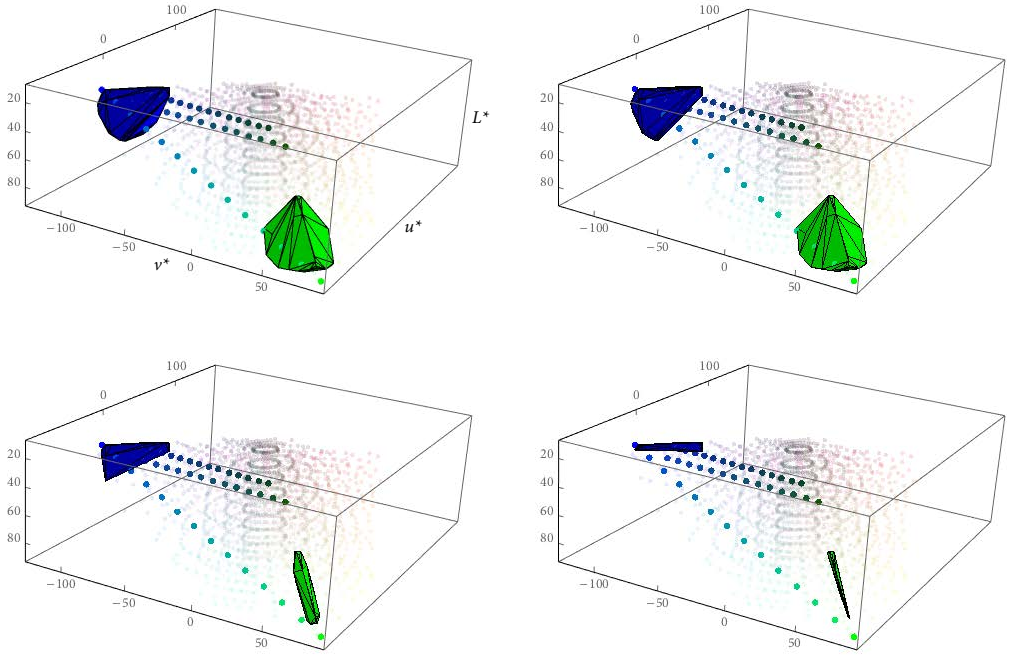

Figure 5.4: CIELUV space with Series 1–3 and prototypical regions for blue and green estimated on the basis of all 2×144 points selected in Experiment 1 (top left); 75 percent highest-rated points (top right); 50 percent highest-rated points (bottom left); 25 percent highest-rated points (bottom right).

In particular, we started by calculating the convex hulls for the sets of points resulting from the choices (i)–(iv); Figure 5.4 shows the convex hulls for these different choices together with Series 1–3. Next, for each of those choices, we sampled 1,000 points from the blue convex hull and 1,000 points from the green convex hull, and we used all 1,000,000 pairs of points $\langle p_b, p_g \rangle$ with p_b coming from the prototypical blue region and p_g coming from the prototypical green region to generate simple Voronoi tessellations on CIELUV space. Note that, because each of these tessellations is generated by two points, it divides CIELUV space into two cells. By the definition of a simple Voronoi tessellation, these cells are separated from each other by the perpendicular bisector plane of the line segment connecting the two generating points. For points $p_1 = (p_{11}, p_{12}, p_{13})$ and $p_2 = (p_{21}, p_{22}, p_{23})$, this perpendicular bisector plane is given by the equation

$$\begin{aligned}(p_{11} - p_{21})x + (p_{12} - p_{22})y + (p_{13} - p_{23})z \;\; &= \\ \frac{1}{2}\langle p_{11} + p_{21}, p_{12} + p_{22}, p_{13} + p_{23}\rangle &\cdot \langle p_{11} - p_{21}, p_{12} - p_{22}, p_{13} - p_{23}\rangle.\end{aligned}$$

One readily verifies that every point in this plane—every point (x, y, z) that satisfies the above equation—is equidistant from p_1 and p_2.

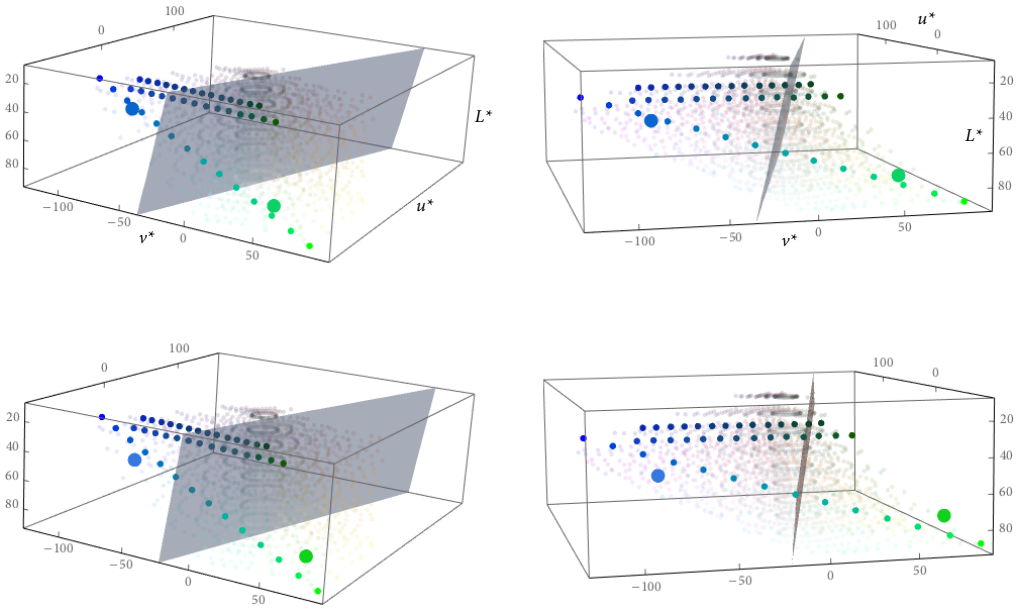

Figure 5.5: Each row shows a different simple Voronoi tessellation on CIELUV space generated by two points randomly chosen from the prototypical blue and prototypical green regions. The pairs of generating points appear bigger than the points representing the series. The blue–green boundary of each tessellation appears as a gray plane.

Figure 5.5 shows two of the 1,000,000 different simple Voronoi tessellations that were generated in this process. It can be seen that, while the tessellation shown in the top row groups patches 1 through 10 of Series 1 with the pick from the blue prototypical region, the tessellation shown in the bottom row groups patches 1 through 11 of Series 1 with that pick. Similarly, the first tessellation groups the first 12 patches of Series 2 and the first 7 patches of Series 3 with the pick from the prototypical blue region while the second groups the first 11 patches of Series 2 and the first 8 patches of Series 3 with the same pick.

The final step, then, consisted of counting, for each of the 42 patches, how many of the 1,000,000 simple Voronoi tessellations grouped it with the pick from the prototypical blue region.[5] Dividing the resulting numbers by 1,000,000 yielded degrees of blueness; degree of greenness was set equal to 1 minus degree of blueness. The top row of Figure 5.6 shows the results of the calculations.

It is obvious from these graphs that the choice of the cutoff point in the first part of the experiment hardly makes a difference, and so it is not surprising that we find perfect correlations between degrees of membership for the same series with different cutoff points for typicality. There are also strong correlations ($r \geqslant .89$) between the degrees for

[5] Practically speaking, this amounted to determining for each of the patches whether it was closer to the pick from the prototypical blue region or to the pick from the prototypical green region.

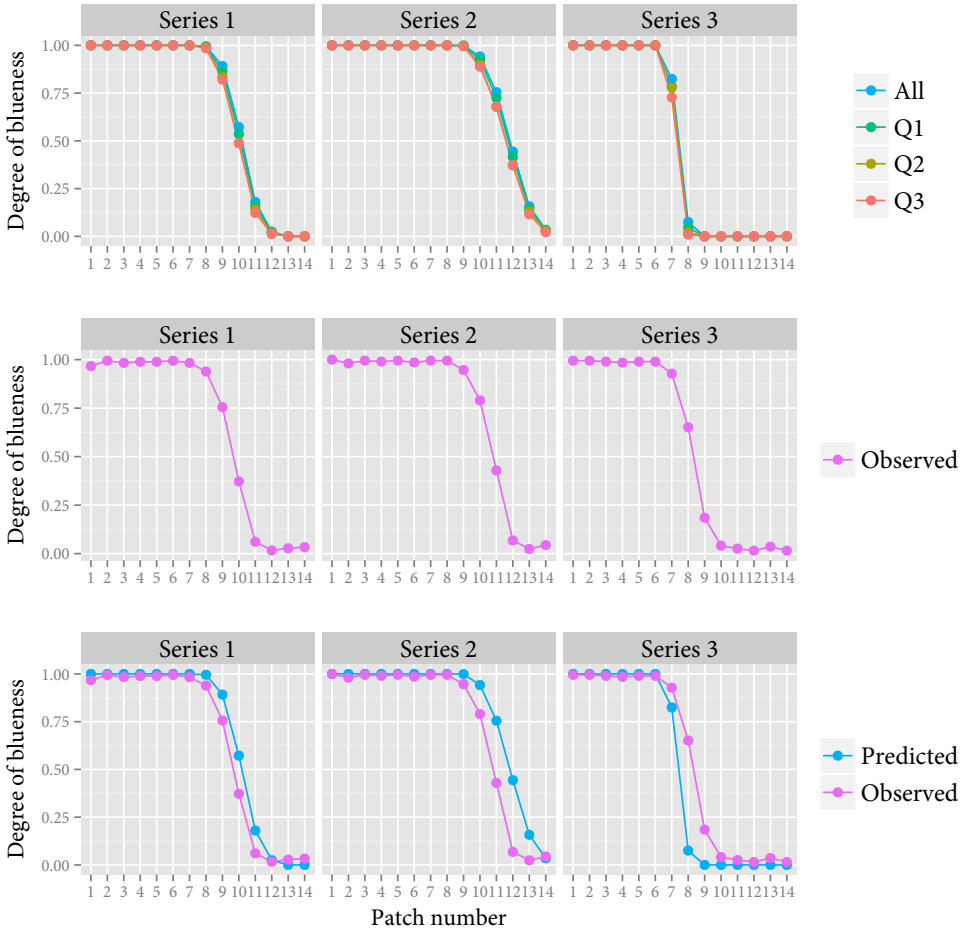

Figure 5.6: Top row: Predicted degrees of blueness for the patches in Series 1–3, assuming different cutoff points for typicality; Q*i* gives the degrees of blueness assuming as prototypical regions the convex hulls of the points with numbers of clicks greater than or equal to the *i*-th quantile. Middle row: Observed degrees of blueness. Bottom row: Predicted (All) versus observed degrees of blueness.

Series 1 and 2, and lower correlations ($.58 \leqslant r \leqslant .78$) between the degrees for either of the first two series and Series 3.

To test Kamp and Partee's account of graded membership, on which the predictions shown in the top row of Figure 5.6 are based, we must confront these predicted degrees of blueness of the various patches in Series 1–3 with the degrees of blueness of those patches as perceived by the participants. In all comparisons of predicted and observed degrees of blueness, we took into account predicted degrees based on each of the aforementioned cutoff points. Given that not only were the qualitative conclusions exactly the same whichever cutoff point was assumed, but even quantitative differences were negligible, we report only analyses assuming predicted degrees as calculated on the basis of the largest

convex hulls.[6] (To allow readers to verify this, the Appendix gives the numerical values of the calculated degrees of membership for all four cutoff points.)

Measuring perceived degrees of blueness was the purpose of the second part of the present experiment. To determine these degrees, we interpreted the proportion of participants who judged a chip to be blue as the degree of blueness of the chip. In doing so, we loosely followed a suggestion made by Hampton (2007, p. 361) in the context of his own work on vagueness and graded membership.[7] The suggestion is that the degree to which an item a is a member of a set S equals the probability that a is classified as belonging to S, where this probability is operationalized in terms of the proportion of people who in fact classify the item as belonging to S. We say that our operationalization of degree of membership is *loosely* based on this suggestion because Hampton considered data obtained from asking whether a given item did or did not belong to a given category. Rather than asking the participants whether or not a color belonged to the blue category, we asked whether it belonged to the green or to the blue category. (More on this methodological issue below.) The middle row of Figure 5.6 shows the proportions-as-degrees for all chips.

5.2.2 Comparing predicted and observed degrees of membership

For a first comparison with predicted degrees, the bottom row of Figure 5.6 plots the observed degrees together with the predicted degrees. We see that there is a very good match between the predictions and the observations for Series 1 as well as a reasonably good match between the predictions and the observations for Series 2. The match between the predictions and observations for Series 3 is not quite as good. What can already be seen from the figure, however, is that, in a clear sense, the predictions for a given series match the observations for that series better than they match the observations for either of the other two series. Notice in particular that that, both for the predictions and for the observations, the Point of Subjective Experience (PSE; here, the patch number corresponding to 50 percent agreement) for Series 1 is to the left of the PSE for Series 2 and to the right of the PSE for Series 3.

To go beyond a mere visual inspection of the data, we fitted logistic curves to both the observed and the predicted degrees for each series, then determined the PSE for each of those curves, and finally calculated the slope of each curve at its PSE. More specifically, for each data series we fitted a logistic curve by estimating values for three parameters in

$$y = \frac{\varphi_1}{1 + \exp[-(x - \varphi_2)/\varphi_3]};$$

[6] Note that it would be a mistake to think that, given that all cutoff points give basically the same results, we might as well have calculated degrees of membership on the basis of prototypical *points* for blue and green. In that case, the collated Voronoi tessellation would be degenerate in that it would really be just a simple Voronoi tessellation. That would mean that the borderline between blue and green is just a plane, which in turn would mean that the graded membership function for blue is a step function, assigning a degree of blueness of 1 to any shade lying on the side of the blue prototype and a degree of blueness of 0 to any shade lying on the other side.

[7] Actually, Hampton attributes the idea to Black (1937). Basically the same suggestion had also been made by the French mathematician Borel in his work on the sorites paradox; see Égré and Barberousse (2014).

Table 5.1: PSEs and slopes for logistic curves fitted to predicted and observed degrees of blueness.

	PSE		Slope	
	Predicted	Observed	Predicted	Observed
S1	10.16 (±0.04)	9.67 (±0.13)	−0.46 (±0.01)	−0.44 (±0.06)
S2	11.82 (±0.05)	10.79 (±0.13)	−0.36 (±0.01)	−0.45 (±0.06)
S3	7.38 (±0.02)	8.30 (±0.11)	−1.02 (±0.03)	−0.51 (±0.09)

we set $y = .5$ for the resulting function, solving for x to obtain the PSE for the given series; and we calculated the first derivative of the function at the PSE, which gave us the slope at that point.[8]

Table 5.1 gives the values of the PSEs and slopes for each series, with 95 percent bootstrap CIs in brackets. For observations, these CIs are based on 1,000 resamplings for each series, but—because of the computational costs of the simulations required to calculate predictions—are based on only 10 resamplings for predictions per series; the latter does not seem to be much of a limitation, given the very low variability in the results, which in fact we had some reason to expect in light of our earlier finding that it made virtually no difference which cutoff point for typicality was chosen to calculate predicted degrees.

The values in Table 5.1 buttress the claim that the predictions match observations better within series than across series: for each series, the PSE for the predicted degrees is closer to the PSE for the observations than it is to the PSE to either of the other series. And while the slopes for the observations for Series 1 and 2 are basically the same (the difference between them is less than 0.005), in contrast to the slopes for their predictions, the slope for the observations for Series 3 is clearly steeper than the slopes for the other two series, in accordance with the predictions. From the CIs, we can easily calculate that these patterns hold reliably.

In a further comparison of predictions and observations, we calculated the sum of squared deviations of the predictions for a given series from the degrees measured for that series, which are given by the diagonal of Table 5.2. Because we were interested in the specificity of our model, for each series we also looked at the sum of squared deviations of the predictions for that series from the observed values for each of the other series, which are also given in Table 5.2. For instance, the entry at the first row, third column, indicates that the sum of squared deviations of the predictions for Series 1 from the observations for Series 3 equals 0.93. More generally, it is seen that, without exception, the predictions for any given series deviate less from the observations for that series than from the observations for either of the other series, in accordance with the findings reported above.

[8] We thank James Hampton and an anonymous referee for suggesting this analysis. As for the details of the analysis, we followed Pinheiro and Bates (2000, Ch. 8).

Table 5.2: Sums of squared deviations of predicted from observed degrees of blueness.

		Observed		
		S1	S2	S3
Predicted	S1	0.08	0.12	0.93
	S2	1.07	0.29	2.33
	S3	1.49	2.56	0.38

5.3 Discussion

This experiment was set up in two parts, the first having the aim of estimating the locations of the prototypical blue and prototypical green regions in CIELUV space—the regions where the typical instances of blue and green are to be found—and the second having the aim of measuring the degrees of blueness/greenness of the various patches in Series 1–3. From the data obtained in the first part in conjunction with the conceptual spaces version of Kamp and Partee's account of graded membership, we could calculate degrees of blueness/greenness for those same patches, which were compared with the degrees of membership attributed by participants.

We obtained very similar S-shaped patterns in the predictions as well as in the observations for all three series. Nevertheless, it was seen that the predictions for a given series invariably fit the observations for that series better than they fit the observations for the other series. It is particularly noteworthy that these findings hold even with respect to Series 1 and 2, which lie very close to one another in CIELUV space. This is evidence that Kamp and Partee's account of graded membership is, or at least can be, highly discriminating.

In addition to this, it is worth highlighting that Kamp and Partee's account of graded membership makes predictions about where in CIELUV space the boundary region between blue and green is to be found, including about where our series of patches enter that region and where they exit it. Specifically, it predicts that, viewed from the side of the prototypical blue region, Series 3 enters the blue/green boundary region earlier than either of the other two series, and Series 1 enters that boundary region a little earlier than Series 2. It further predicts that Series 3 also exits the boundary region earlier than the other two series, and Series 1 exits it a little earlier than Series 2. Figure 5.6 makes it easy to see that all these predictions are borne out by the data.

Most importantly, we obtain all this—moderately good to very good matches between predicted and observed degrees of membership, as well as accurate qualitative predictions concerning the blue/green boundary region—*without having estimated any free parameters*. To the contrary, all predictions followed ultimately from a semantics the development of which was mainly driven by linguistic and philosophical concerns. That nevertheless the semantics is at least fairly successful in predicting the data obtained in Experiment 2 points toward its correctness. Indeed, we are not aware of any psychological

model of categorization that can presently offer quite the same. See the general discussion for more on this.

While, as we said, the semantics is fairly successful in predicting the data, we saw that in particular the match between the predictions and the observations for Series 3 leaves something to be desired. In this connection, it is worth recalling some of the limitations of our studies which we acknowledged in Section 3. That our data were obtained via online surveys rather than in a specialized color laboratory, and that we are assuming CIELUV space, which is known to be not entirely satisfactory as a model of human color perception—if indeed there can be one such model (see Hardin, 1999)—already implies that our predictions can be at best good approximations of the data. In fact, with respect to the color space we are assuming, it may be especially noteworthy that Series 3, for which our predictions were not as good as for Series 1 and 2, has a span in CIELUV space that is roughly twice that of the other series (229 versus 123 for Series 1 and 102 for Series 2). Color scientists have suggested that color space may actually be only *locally* Euclidean (Indow, 1988), which has led some to propose different metrics for measuring small and large distances in color space (Westland, Ripamonti, & Cheung, 2012, p. 68 ff).[9] This may make it more problematic to conceive—as we have been doing—of Series 3 as one series of equally spaced patches in Euclidean space than it is to conceive of the other two series as such. It may also mean that our model needs some slight modification. For instance, as the model stands, all simple Voronoi tessellations contribute equally to degree of membership, in that the degree of membership of item i in category C is given by the measure of simple Voronoi tessellations that group i with the prototypical C region. But one could consider *weighting* the individual tessellations differently, with the weights being some function of the distances between the generating points.[10] This might require adding one or more parameters to the model. Whether this, or something along these lines, would significantly improve model fit in the case at hand is a question we relegate to future work. For now, we content ourselves with noting that even in its present form, without any free parameters, Kamp and Partee's semantics does, overall, at least reasonably well in light of the data gathered in this experiment.

6 Experiment 3

In the second part of Experiment 2, we used a two-alternative forced choice task, following a suggestion of Hampton's. But there are other plausible procedures for measuring the degrees of blueness/greenness for the 42 patches from Series 1–3. To make our conclusions as robust as possible, and to control for the possibility that the results obtained in Experiment 2 might be a bias of the 2AFC task employed, we followed up with Experiment 3 in which we measured degrees of membership in a number of different ways. The

[9] Alternatively, one could consider making color similarity an exponentially decaying function of Euclidean distance in CIELUV space; see, e.g., Nosofsky (1986, 1987), and Shepard (1986).

[10] We here mean the possibility of letting the individual tessellations weigh differently in the calculation of degrees of membership. However, one could also consider weighting the generating points of the individual tessellations differently; see Okabe et al. (2000, Ch. 3) for several ways of doing this.

goal was to compare the outcomes from the different procedures with one another as well as—most importantly—with the predictions obtained from the first part of Experiment 2.

6.1 Method

Participants

There were 357 participants in this experiment, all from Australia, Canada, Great Britain, or the United States. The participants were recruited and tested in the same way as in the previous experiments. They received a small amount of money for their cooperation.

Excluded from the analysis were 6 participants who returned incomplete response sets, 7 participants who indicated that they were colorblind, 21 nonnative speakers of English (the language of the survey), and the fastest and slowest 5 percent of responders.[11] This left us with 288 participants. Further excluded were 33 participants who failed the same sorting task that had been used in the previous experiments and one participant who answered in the negative the question of whether he or she had responded seriously, leaving 254 participants for the final analysis.

These participants spent on average 450 seconds on the survey (SD = 140 s). Their mean age was 39 (SD = 12). Fifty-nine percent of the participants were female. Sixty-eight percent indicated university as their highest education level, 28 percent high school, and the remaining 4 percent a lower education level. Participants included in the analysis did not differ significantly from excluded participants in age, gender, or level of education.

Materials and procedure

The participants were randomly assigned to one of four groups. All four groups were asked to judge the blueness/greenness of all patches in Series 1–3, though for each series, the questions were different, and which series went with which type of questions differed per group. Specifically, we used four types of questions for eliciting degrees of blueness/greenness. Type I was a repetition of the 2AFC task used in the previous experiment; so, participants were given the options "Blue" and "Green" and asked to select the option that described the shown color best. In Type II, participants could move a slider between 0 and 100, where 0 was labeled "Green" and 100 was labeled "Blue"; they were asked where, on the line between "Green" and "Blue," they would locate the color that was shown. In Type III, participants were again given a slider task but were now asked to indicate how blue they deemed the shown color patch, noting that 0 stood for being clearly not blue and 100 for being clearly blue. Type IV was exactly like Type III except that Type IV questions asked for how green the patch was instead of how blue.

As stated, each of the groups was shown all three series. Each chip was shown individually on the screen. The order in which the series appeared was randomized per participant, as was the order in which the patches in the series appeared. The first group (N = 58) was shown Series 1 in combination with Type I questions ("blue or green"), Series 2 in combination with Type II questions ("where between green and blue . . . "), and Series 3 in combination with Type III questions ("how blue . . . "); the second group

[11] Here, too, all analyses were repeated with fastest and slowest responders included, and again this did not lead to any qualitatively different results.

(N = 61) was shown Series 1 in combination with Type IV questions ("how green . . . "), Series 2 in combination with Type I questions, and Series 3 in combination with Type II questions; the third group (N = 67) was shown Series 1 in combination with Type III questions, Series 2 in combination with Type IV questions, and Series 3 in combination with Type I questions; and finally the fourth group (N = 68) was shown Series 1 in combination with Type II questions, Series 2 in combination with Type III questions, and Series 3 in combination with Type IV questions.

6.2 Results

The top row of Figure 6.1 gives a graphical summary of the data, where data obtained by means of Type I questions represent proportions of "Blue" responses and all other data have been scaled by dividing them by 100; in addition, data obtained by means of Type IV questions, which asked for degrees of greenness, have, for reasons of comparability, been first divided by 100 and then subtracted from 1 (so we obtained degrees of blueness from degrees of greenness). To facilitate comparison with the observed degrees from Experiment 2, these are also plotted in the graphs in the top row of Figure 6.1; to facilitate comparison with the predicted degrees, these are shown again in the bottom row.

It is readily seen that the degrees of blueness obtained via the questions of Types II–IV all form somewhat flatter S-shapes than the degrees obtained via Type I questions (including the degrees observed in the previous experiment). Visually, the former group of measurements deviate more from the predicted degrees of blueness than the latter

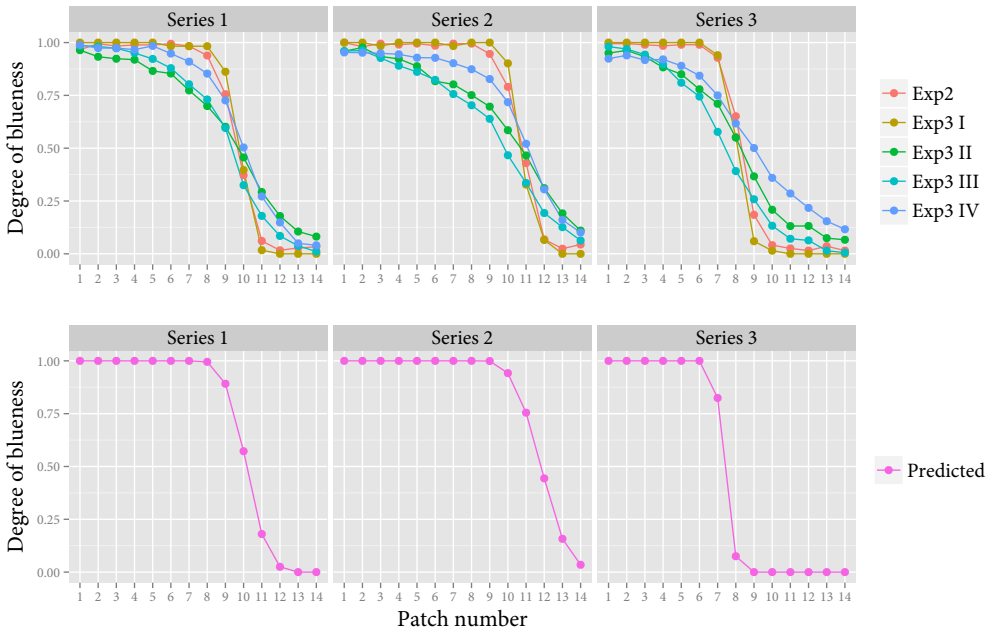

Figure 6.1: Observed (top row) and predicted (bottom row) degrees of blueness.

Table 6.1: PSEs and slopes for logistic curves fitted to observed degrees of blueness as measured in Experiment 3.

	Type I		Type II	
	PSE	Slope	PSE	Slope
S1	9.72 (±0.20)	−0.52 (±0.18)	10.76 (±0.21)	−0.47 (±0.02)
S2	10.76 (±0.17)	−0.47 (±0.23)	10.25 (±0.21)	−0.10 (±0.02)
S3	8.06 (±0.16)	−0.69 (±0.20)	8.12 (±0.19)	−0.13 (±0.04)
	Type III		Type IV	
	PSE	Slope	PSE	Slope
S1	8.98 (±0.20)	−0.17 (±0.03)	9.87 (±0.19)	−0.18 (±0.03)
S2	9.38 (±0.33)	−0.11 (±0.02)	10.87 (±0.20)	−0.12 (±0.04)
S3	7.35 (±0.25)	−0.17 (±0.02)	9.10 (±0.38)	−0.11 (±0.02)

group. This impression is confirmed by looking at the PSEs and slopes of the logistic curves that were fitted to each series in the way described in Section 5.2.2 and which are displayed in Table 6.1 (in brackets, bootstrap CIs based on 1,000 resamplings of the data).

It will be remembered that, for the predictions, the PSE for Series 1 was reliably between those for Series 2 and 3, with the PSE for Series 3 being the lowest. We found the exact same pattern in the observations from Experiment 2 and find it here as well, with one exception: for the degrees obtained via the Type II questions, the PSE for Series 2 is between the PSEs for the other two series. The slopes are rather unspecific here, which was already obvious from Figure 6.1.

Here, too, we calculated sums of squared deviations of the predictions for a given series from the observations for that series, which are given by the diagonal of Table 6.2. For reasons explained earlier, the same table gives sums of squared deviations of the predictions for a given series and the observations for the *other* series, showing that, within any question type, predictions for a given series are closer to the observations for that series than they are to the observations for either of the other series.

6.3 Discussion

The purpose of this experiment was to determine to what extent the outcomes of the previous experiment depended on the particular method for measuring degrees of membership that was used in that experiment. In Experiment 2, we measured degrees of blueness indirectly, by making participants choose between "Blue" and "Green" and then equating degree of blueness with the proportion of participants who had chosen the "Blue" option. In the present experiment, we also directly asked people to indicate degree of blueness (greenness). We compared the degrees thus obtained with the predictions based on the first part of Experiment 2.

Table 6.2: Sums of squared deviations of predictions from observations.

		Observed					
		Type I			Type II		
		S1	S2	S3	S1	S2	S3
Predicted	S1	0.06	0.14	1.24	0.35	0.41	0.81
	S2	1.08	0.35	2.74	0.88	0.49	1.81
	S3	1.75	2.81	0.24	1.16	1.71	0.56
		Observed					
		Type III			Type IV		
		S1	S2	S3	S1	S2	S3
Predicted	S1	0.29	0.36	1.27	0.09	0.30	0.55
	S2	1.15	0.81	2.50	0.63	0.20	1.13
	S3	0.96	1.27	0.37	1.50	2.27	0.91

As in Experiment 2, we found that total squared deviations were smaller between predictions for a given series and observations for that series than between those predictions and the observations for the other series. Similarly, we again found that almost without exception the PSE for Series 1 lay between those for the other series, as predicted. These findings attest further to the accuracy of the account of graded membership at issue.

On the other hand, the match between the predictions and the observations obtained by directly asking for degrees of membership was less good than the match between the same predictions and the observations obtained via the 2AFC task used in the previous experiment and in the 2AFC task (the Type I questions) in the present experiment. This difference in the results was unanticipated but is, we believe, not hard to explain.

First, there is a well-documented bias—the central tendency bias—that inclines participants to avoid the endpoints of a response scale. According to Stevens (1971, p. 428), this is "one of the most obstinate" response biases, which makes it reasonable to believe that it contributed to the fact that, in the responses to the slider questions (the Type II–IV questions), even those patches in Series 1–3 that one would expect to have appeared clearly blue or clearly green to participants receive non-extreme degrees. Obviously, the 2AFC tasks that we also used cannot have been affected by this bias.

More importantly, there is a key respect in which measuring degrees of blueness via the 2AFC task is exactly parallel to predicting degrees of blueness on the basis of the prototypical blue and green regions. If the patches of Series 1–3 had been presented to participants in a free naming task, then very probably next to blue and green responses there would have been purple, turquoise, and possibly also gray responses; at any rate, it is clear that there are patches for which "blue" or "green" are not necessarily the best or only plausible labels. However, these were the only labels available to the participants in

Experiment 2 as well as to the participants in the first group of Experiment 3. Now note that it is also highly probable that, if we had been able to take into account information about the locations in CIELUV space of the prototypical regions of purple and gray, that would have influenced the outcomes of our predictions of degrees of blueness. In our calculations of degrees of membership, each simple Voronoi tessellation classified a patch either as blue or as green. If these calculations could have been based on knowledge of the locations of additional prototypical regions, so that some simple Voronoi tessellations might have classified a patch as purple or gray, some patches might have been predicted to be purple to some degree, or gray to some degree. In other words, our predictions were, in a way, also based on a 2AFC task.

Finally, in Experiment 2 and Experiment 3, Type I group, where we used a 2AFC task, one expects for a single participant (and not taking into account errors or inconsistencies in the responses) the patches in a series to be judged blue until a certain patch, and green afterwards. In other words, one essentially expects a step function. It is also reasonable to expect the "switch point" to vary across participants, such that averaging over the responses will yield an S-shaped curve—which is the result we obtained. By contrast, in Experiment 3, Type II–IV groups, where we used a slider task, one expects for a single, perfectly consistent participant an S-shaped response curve. Moreover, for Experiment 3, Type I group (which was presented the most neutral phrasing of the slider task), and assuming an "ideal" participant, one would expect the PSE to occur between the same patches as the switch point in the 2AFC task. Hence, averaging over the responses of the slider tasks yields an S-shaped curve due to *two* factors: intra-participant (individual graded response) and inter-participant (variation as before). In tandem, these factors cause the slopes for the slider questions to be lower than the slopes of the S-curves resulting from the 2AFC tasks—which is again what we find.[12]

7 General discussion

Linguists and logicians have expended much time and effort on constructing formal models of language. It is hard to overestimate the importance of this work for our understanding of computer languages as well as the language of mathematics. But because most of this work implicitly assumes predicates to be crisp, in that they either apply or do not apply to any given item, formal semantics has improved our understanding only for very limited fragments of *natural* languages, for it is a characteristic feature of such languages that many of their predicates are vague and can apply to different items in differing partial degrees.

Kamp and Partee have made important progress toward constructing a formal semantics that can handle such vagueness of meaning.[13] By their own admission, the account of the graded membership relation that is at the heart of their semantics was still unfinished.

[12] Thanks to an anonymous referee for pressing us on this.

[13] As Kamp and Partee acknowledge in their 1995 paper, in constructing their semantics they could build on important work in fuzzy logic (e.g., Zadeh, 1995). However, a crucial difference between their work and the earlier work in fuzzy logic is that the latter took a notion of graded membership for granted whereas Kamp and Partee aimed to define it.

However, later work by other researchers showed how that account could be completed in a natural way in the conceptual spaces framework that had been developed in the cognitive sciences for purposes not primarily related to vagueness. The thus completed semantics could also be shown to be formally correct in the sense that it does not give rise to conflicting semantic assessments of one and the same sentence.

As mentioned in the introduction, formal correctness is one of two broadly accepted desiderata for semantics, the other one being material adequacy. This paper reported the results of a number of experiments conducted to test the material adequacy of Kamp and Partee's semantics in the conceptual spaces version. We obtained estimates of the locations of the prototypical blue and prototypical green regions in CIELUV space, from which, via Kamp and Partee's account, degrees of blueness/greenness for various shades in the blue–green region of color space could be derived. These were compared to the second part of our results, which consisted of measurements of degrees of blueness.

The results of this comparison spoke rather unambiguously in favor of Kamp and Partee's account, in that the predictions were largely borne out by the observations. The upshot of our experiments was not just that, overall, the predictions matched the observations quite well, but also that Kamp and Partee's account turned out to be highly discriminative in that its predictions for Series 1 matched more closely the observations for that series than the observations for the nearby Series 2 (nearby in CIELUV space, that is), and vice versa.

As a formal semantics for languages with vague predicates, Kamp and Partee's account has few competitors. But there are psychological accounts of categorization that can model vagueness and that could perhaps serve as a basis for a formal semantics. Therefore, one might regard these accounts as being in competition with Kamp and Partee's. To bring further into relief the virtues of the latter account, we briefly compare it to two prominent models of categorization that allow for vague categories, namely, Nosofsky's (1986, 1987, 1989) Generalized Context Model (GCM) and Hampton's (2007) Threshold Model (TM).[14]

Nosofsky's model assumes exemplar theory rather than prototype theory and aims to predict categorization probabilities rather than degrees of membership. Also, as mentioned in note 9, Nosofsky assumes similarity to be measured by an exponential decay function. But one can make the GCM at least to some extent comparable to Kamp and Partee's account by interpreting categorization probabilities as degrees of membership and by agreeing to measure color similarity differently than is standardly done in CIELUV space; for reasons given earlier in the paper, neither proposal would be unreasonable.

On the GCM, we learn a category by committing to memory all examples of the category that we have come across so far. The probability that a newly encountered item is included in the category then depends on how similar that item is to the stored exemplars. In more detail, where we have two categories, C_1 and C_2, the probability that an item i is

[14]See also Verheyen, Hampton, and Storms (2010), which presents the Rasch model (Rasch, 1960) as a formalization of the TM.

classified as belonging to C_1 is given by

$$\frac{\beta\,\mathrm{sim}(i, C_1)}{\beta\,\mathrm{sim}(i, C_1) \;+\; (1-\beta)\,\mathrm{sim}(i, C_2)},$$

with β ($0 \leqslant \beta \leqslant 1$) representing the response bias for category C_1. This equation has the same form as Medin and Schaffer's (1978) Context Model, but it is more general than that because of the way $\mathrm{sim}(i, C_n)$, the sum of the similarities of i to the exemplars of C_n ($n = 1, 2$), is defined. Specifically, $\mathrm{sim}(i, C_n) \coloneqq \sum_j \big(\exp(-\lambda\,\mathrm{dist}(i, j)^p\big)$, where λ is the typicality gradient, which determines the steepness of the exponential decay; the choice for p depends on the discriminability of the stimuli (Shepard, 1986); and $\mathrm{dist}(i, j)$ is defined as $\big(\sum_{k=1}^{m} w_k |x_{ik} - x_{jk}|^r\big)^{1/r}$ for an m-dimensional space. In the definition of dist, the w_k's take values in $[0, 1]$ and add up to 1; w_k is the attention weight given to the k-th dimension of the space in computing the overall distance, and x_{ik} is item i's k-coordinate in that space.

We could let the prototypical points sampled in Experiment 1 (or some subset thereof) be proxy for exemplars and then try to find values for the parameters β, λ, p, r, w_1, w_2, and w_3 that jointly minimize the sum of squared deviations of classification probabilities from measured degrees of membership for the three series of color patches that figured in our experiments, or which maximize the correlations between classification probabilities and measured degrees of membership, or which strike the best balance between doing both.[15] Rather than tackle this optimization problem, we note that Kamp and Partee's account presents us with no such task. As was emphasized before, degrees of membership flow directly from that account in conjunction with the structure of color space (including the locations of prototypical regions), and so no model fitting is involved. That makes Kamp and Partee's account predictively much more specific than the GCM: instead of asserting that values can be found for certain parameters which will yield degrees of membership that match the observations, it directly predicts these degrees of membership. And with all the caveats that apply in light of the background assumptions (such as the assumption of CIELUV space and the assumption that we can average over participants' judgments), the match between predictions and observations is quite accurate. This is a virtue that will stand even if it turns out that the GCM yields a more accurate fit—which is by no means certain it can, although with seven parameters to be fit, it would also not be hugely surprising.[16]

A comparison of Kamp and Partee's account with Hampton's TM leads to essentially the same conclusion. Hampton defines graded membership as a function of similarity to category prototype. He assumes that each concept has a determinate boundary region for

[15] This was suggested to us by James Hampton, for which we thank him.

[16] In fairness to the GCM, some parameters will be fixed by the choice of stimuli on independent (theoretical or empirical) grounds. For instance, for stimuli with integral dimensions (such as color stimuli) it is commonly assumed that 2 is the appropriate value for the parameter r in the definition of the dist function ($r = 2$ turns dist into a weighted Euclidean distance function). On the other hand, it is to be noted that later versions of the GCM have parameters in addition to the ones mentioned here, like a response-scaling parameter which appears as an exponent of the summed similarities in the formula of the generalization of Medin and Schaffer's Context Model displayed in the text; see Ashby and Maddox (1993), Nosofsky and Zaki (2002), and Navarro (2007).

membership; in his notation, S_L and S_H are the values of the relevant similarity measure S that constitute the lower and upper bound, respectively, of the boundary region of a given category. S_T denotes the value at which the degree of membership in that category is .5; this is supposed to be midway between S_L and S_H. Then the degree of membership of an item i in the category, $M(i)$, is defined thus:[17]

$$M(i) = \begin{cases} 0 & \text{if } S_L \geqslant S(i); \\ 2\left(\frac{S(i)-S_L}{S_H-S_L}\right)^2 & \text{if } S_T \geqslant S(i) > S_L; \\ 1-2\left(\frac{S_H-S(i)}{S_H-S_L}\right)^2 & \text{if } S_H \geqslant S(i) > S_T; \\ 1 & \text{if } S(i) > S_H. \end{cases}$$

Hampton highlights various attractive features of this proposal. For instance, it accords with pretheoretical intuition that there are regions of determinate membership and determinate non-membership where the membership function takes the values 1 and 0, respectively, and the definition is also consistent with the more than plausible assumption that items whose degree of membership is 1 can still vary in the degree to which they are *typical* for the category. Moreover, it provides a straightforward solution to the problem of second-order vagueness, that is, the fact that we do not sense sharp dividing lines, on the one hand, between the clear and the not-so-clear members of a category and, on the other hand, between the not-so-clear members and the clear non-members (Sainsbury, 1991; Keefe & Smith, 1997, p. 15 f). To explain this, it is enough to point out that, on the current proposal, the transitions at the endpoints of the boundary region are perfectly smooth, so that we cannot expect to experience any abruptness in going from the clear members to the not-so-clear members (or the other way round) or from the not-so-clear members to the clear non-members (or the other way round).

The TM was an important source of inspiration for the version of Kamp and Partee's account presented in Decock and Douven (2014). Among other things, these authors adopted Hampton's treatment of the issue of second-order vagueness. As they also point out, however, while the TM at first appears very specific—what justifies the particular choice of sigmoid function made in the definition of M, or the suggestion that the sigmoid function would be the same for every category?—it is actually rather *un*specific, given that, in a note, Hampton (2007, p. 381 n) distances himself from the particular definition of graded membership given by M, saying that it is meant only to serve illustrative purposes. For our experiments, this means that the TM predicts no more than that degrees of blueness will be 1 for any patch closer to the blue prototype than S_H, 0 for any patch further away from the blue prototype than S_L, and for the patches in between will decline in some S-shaped pattern as they are further removed from the blue prototype. Clearly, that is what the conceptual spaces version of Kamp and Partee's semantics predicts too (with "prototypical blue region" replacing "blue prototype"). In addition, however, that account predicts exactly where the boundary region begins and where it ends, as well as what the S-shape in between looks like. If one is free to pick points S_L and S_H, and

[17] In Hampton's definition, there occurs a > wherever we have a ⩾. That must be an oversight, as it would make M undefined at S_L, S_T, and S_H, which cannot be intended.

free to choose any sigmoid function connecting those points in *something like* the way M does, then one will almost surely be able to obtain better fitting models than the ones provided by Kamp and Partee's account. But it is hard to see how this could be a point in favor of the TM. Otherwise, why not settle on a theory according to which there is *some* function—whether or not sigmoid—that describes the decline of membership in the boundary region of a category? With virtually *no* constraints on the shape of the function, we will certainly be able to get still better fitting models.

In short, there is a clear sense in which Kamp and Partee's account is preferable to both the GCM and the TM, to wit, it is more specific than either. At the same time, its surplus empirical content makes it more vulnerable to refutation. We saw, however, that the account still does quite well in light of the data gathered in our experiments. This is important in its own right, regardless of whether extant psychological accounts of categorization are able to explain the same data equally well (or even better). For Kamp and Partee's account offers something in addition to an account of categorization, being first and foremost a formal device that lets us compute truth values of sentences of arbitrary syntactic complexity in a systematic manner, where the sentences may contain vague predicates. No psychological account of categorization offers anything close to that, and none of the rival semantics developed by linguists and philosophers is formally as elegant or empirically as powerful as Kamp and Partee's.

The experiments presented in the current study were limited to two vague predicates pertaining to the color domain. We acknowledged that this may have influenced our results. This admission suggests obvious follow-up research, namely, to rerun Experiment 1 and the first part of Experiment 2 for other colors, perhaps for all of Berlin and Kay's nine other basic color categories, and then to calculate anew degrees of membership for the patches in Series 1–3. Indeed, a more general extension of the present study would also consider series of color patches located in other regions than the blue–green region.

We also mentioned various ways in which our design could be refined. Most notably, we expect that repeating basically the same experiments under better controlled testing conditions and testing fully within participants—meaning that color similarity space, locations of prototypical regions in that space, and judgments of degrees of membership for various colors, are all determined per participant rather than at the aggregate level—would lead to a better fit between predictions and observations. We also considered the possibility of tweaking the conceptual spaces version of Kamp and Partee's semantics by assigning different weights to the tessellations that make up a collated Voronoi diagram; that way, different tessellations might make different contributions to determining degrees of membership.

Finally, given that we cannot assume that what holds for the domain of color will hold for other domains of concepts as well, the follow-up research most urgently called for is testing Kamp and Partee's account using conceptual spaces other than color space. A number of candidate spaces have been mentioned at the end of Section 2. While in testing Kamp and Partee's account in other domains one is likely to encounter difficulties specific to that domain,[18] we strongly suspect that the overall design of our experimental

[18] For instance, an anonymous referee noted that other sets of visual stimuli, such as facial expressions of

work can to a large extent serve as a template for such future work. And we are even more strongly convinced that the results of the current experiments are encouraging enough to warrant embarking on that work.

Acknowledgments

We are greatly indebted to three anonymous referees for extensive and valuable comments and to James Hampton for very helpful editorial comments. We are further grateful to Danny Vanpoucke for his help with creating the stimuli for Experiment 1 and to Christopher von Bülow and Frank Zenker for useful feedback.

References

Ashby, F. G. and Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37*, 372–400.

Aust, F., Diedenhofen, B., Ullrich, S., and Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods, 45* 527–535.

Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101–140). Cambridge: Cambridge University Press.

Benavente, R., Vanrell, M., and Baldrich, R. (2006). A data set for fuzzy colour naming. *Color Research and Application, 31*, 48–56.

Berlin, B. and Kay, P. (1969/1999). *Basic Color Terms*. Stanford CA: CSLI Publications.

Black, M. (1937). Vagueness: An exercise in logical analysis. *Philosophy of Science, 4*, 427–455.

Borg, I. and Groenen, P. (2010). *Modern multidimensional scaling* (2nd ed.). New York: Springer.

Bosten, J. M., Robinson, J. D., Jordan, G., and Mollon, J. D. (2005). Multidimensional scaling reveals a color dimension unique to "color-deficient" observers. *Current Biology, 15*, R950–R952.

Castro, J. B., Ramanathan, A., and Chennubhotla, C. S. (2013). Categorical dimensions of human odor descriptor space revealed by non-negative matrix factorization. *PLoS ONE*, 8:e73289, doi: 10.1371/journal.pone.0073289.

Churchland, P. M. (2012). *Plato's camera*. Cambridge MA: MIT Press.

Clark, A. (1993). *Sensory qualities*. Oxford: Clarendon Press.

Cook, R. S., Kay, P., and Regier, T. (2005). The World Color Survey database: History and use. In H. Cohen and C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 223–242). Amsterdam: Elsevier.

emotion, may pose specific problems given that, unlike in the color domain, one cannot vary the stimuli along one dimension while holding the values for the other dimensions fixed.

Decock, L. and Douven, I. (2011). Similarity after Goodman. *Review of Philosophy and Psychology, 2*, 61–75.
Decock, L. and Douven, I. (2014). What is graded membership? *Noûs, 48*, 653–682.
Douven, I. and Decock, L. (2016). What verities may be. *Mind*, in press.
Douven, I., Decock, L., Dietz, R., and Égré, P. (2013). Vagueness: A conceptual spaces approach. *Journal of Philosophical Logic, 42*, 137–160.
Égré, P. and Barberousse, A. (2014). Borel on the heap. *Erkenntnis, 79*, 1043–1079.
Fairchild, M. D. (2013). *Color appearance models*. Hoboken NJ: Wiley.
Gärdenfors, P. (2000). *Conceptual spaces*. Cambridge MA: MIT Press.
Gärdenfors, P. (2007). Representing Actions and functional properties in conceptual spaces. In T. Ziemke, J. Zlatev, and R. M. Frank (Eds.), *Body, language and mind* (Vol. 1, pp. 167–195). Berlin: De Gruyter.
Gärdenfors, P. (2014). *The geometry of meaning*. Cambridge MA: MIT Press.
Gärdenfors, P. and Warglien, M. (2012). Using concept spaces to model actions and events. *Journal of Semantics, 29*, 487–519.
Gärdenfors, P. and Zenker, F. (2011). Using conceptual spaces to model the dynamics of empirical theories. In E. J. Olsson and S. Enqvist (Eds.), *Belief revision meets philosophy of science* (pp. 137–153). New York: Springer.
Gärdenfors, P. and Zenker, F. (2013). Theory change as dimensional change: Conceptual spaces applied to the dynamics of empirical theories. *Synthese, 190*, 1039–1058.
Halmos, P. R. (1974). *Measure theory*. New York: Springer.
Hampton, J. A. (1995). Testing the prototype theory of concepts. *Journal of Memory and Language, 34*, 686–708.
Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science, 31*, 355–384.
Hardin, C. L. (1999). Color relations and the power of complexity. *Behavioral and Brain Sciences, 22*, 953–954.
Helm, C. E. (1964). A multidimensional ratio scaling analysis of perceived color relations. *Journal of the Optical Society of America, 54*, 256–262.
Indow, T. (1988). Multidimensional studies of Munsell color solid. *Psychological Review, 95*, 456–470.
Kamp, H. and Partee, B. (1995). Prototype theory and compositionality. *Cognition, 57*, 129–191.
Keefe, R. and Smith, P. (1997). Introduction: Theories of vagueness. In R. Keefe and P. Smith (Eds.), *Vagueness: A reader* (pp. 1–57). Cambridge MA: MIT Press.
Malacara, D. (2002). *Color vision and colorimetry: Theory and applications*. Bellingham WA: SPIE Press.
Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.
Moroney, N. (2003). Unconstrained web-based color naming experiment. *Color imaging VIII: Processing, hardcopy, and applications* (Proc. SPIE, vol. 5008), pp. 36–46.
Mylonas, D. and MacDonald, L. (2010). Online colour naming experiment using Munsell samples. *Proceedings of the 5th European conference on colour in graphics, imaging, and vision*, pp. 27–32.

Mylonas, D., Paramei, G. V., and MacDonald, L. (2014). Gender differences in colour naming. In W. Anderson, C. P. Biggam, C. Hough, and C. Kay (Eds.), *Colour studies: A broad spectrum* (pp. 225–239). Philadelphia PA: John Benjamins Publishing Company.

Navarro, D. J. (2007). On the interaction between exemplar-based concepts and a response scaling process. *Journal of Mathematical Psychology, 51*, 85–98.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 87–108.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception and Psychophysics, 45*, 279–290.

Nosofsky, R. M. and Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 924–940.

Oddie, G. (2005). *Value, reality, and desire*. Oxford: Oxford University Press.

Okabe, A., Boots, B., Sugihara, K., and Chiu, S. N. (2000). *Spatial tessellations* (2nd ed.). New York: Wiley.

Osherson, D. N. and Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition, 29*, 259–288.

Paramei, G. V., Izmailov, C. A., and Sokolov, E. N. (1991). Multidimensional scaling of large chromatic differences by normal and color-deficient subjects. *Psychological Science, 2*, 249–259.

Petitot, J. (1989). Morphodynamics and the categorical perception of phonological units. *Theoretical Linguistics, 15*, 25–71.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed effects models in S and S-Plus*. New York: Springer.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Regier, T., Kay, P., and Cook, R. (2005). Focal colors are universal after all. *Proceedings of the National Academy of Sciences, 102*, 8386–8391.

Rosch, E. (1973). Natural categories. *Cognitive Psychology, 4*, 328–350.

Rosch, E. (1978). Principles of categorization. In E. Rosch and B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale NJ: Erlbaum.

Rosch, E. and Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605.

Sainsbury, M. (1991). Is there higher-order vagueness? *Philosophical Quarterly, 41*, 167–182.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1*, 54–87.

Shepard, R. N. (1986). Discrimination and generalization in identification and classification: Comment on Nosofsky. *Journal of Experimental Psychology: General, 115*, 58–61.

Shepard, R. N. and Carroll, J. D. (1966). Parametric representation of nonlinear data structures. In P. R. Krishnayah (Ed.), *Multivariate analysis* (pp. 561–592). New York: Academic Press.

Shepard, R. N. and Cooper, L. A. (1992). Representations of colors in the blind, color-blind, and normally sighted. *Psychological Science, 3*, 97–104.

Sivik, L. and Taft, C. (1994). Color naming: A mapping in the IMCS of common color terms. *Scandinavian Journal of Psychology, 35*, 144–164.

Sprow, I., Barańczuk, Z., Stamm, T., and Zolliker, P. (2009). Web-based psychometric evaluation of image quality. *Image quality and system performance VI* (Proc. SPIE, vol. 7242), pp. 1–12.

Stevens, S. S. (1971). Issues in psychophysical measurement. *Psychological Review, 78*, 426–450.

Sturges, J. and Whitfield, T. W. A. (1995). Locating basic colours in the Munsell space. *Color Research and Application, 20*, 364–376.

Tarski, A. (1935). Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica, 1*, 261–405.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.

Verheyen, S., Hampton, J. A., and Storms, G. (2010). A probabilistic threshold model: Analyzing semantic categorization data with the Rasch model. *Acta Psychologica, 135*, 216–225.

Wang, J., Green, J. R., Samal, A., and Yunusova, Y. (2013). Articulatory distinctiveness of vowels and consonants: A data-driven approach. *Journal of Speech, Language, and Hearing Research, 56*, 1539–1551.

Westland, S., Ripamonti, C., and Cheung, V. (2012). *Computational colour science using MATLAB* (2nd ed.). Chichester UK: Wiley.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*, 338–353.

Appendix

In Experiment 2, degrees of blueness for the patches in the three color series were calculated on the basis of different cutoff points for which points in CIELUV space were to be included in the prototypical blue and green regions. The analyses presented in the main part of the paper are all based on the degrees of blueness obtained by assuming the most inclusive cutoff point. It was claimed that this choice made no relevant difference. Interested readers can verify this claim by repeating the analyses with the predicted degrees of blueness obtained on the basis of the other cutoff points, which are included in Table A.

Table A: Degrees of blueness for the patches in Series 1–3, assuming different cutoff points for typicality.

	All			Q1			Q2			Q3		
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7	1.00	1.00	0.82	1.00	1.00	0.73	1.00	1.00	0.78	1.00	1.00	0.78
8	1.00	1.00	0.08	0.98	1.00	0.01	0.99	1.00	0.02	0.99	1.00	0.05
9	0.89	1.00	0.00	0.82	0.99	0.00	0.83	1.00	0.00	0.87	1.00	0.00
10	0.57	0.94	0.00	0.49	0.89	0.00	0.49	0.90	0.00	0.54	0.92	0.00
11	0.18	0.75	0.00	0.12	0.68	0.00	0.14	0.68	0.00	0.16	0.73	0.00
12	0.03	0.44	0.00	0.01	0.37	0.00	0.02	0.37	0.00	0.02	0.42	0.00
13	0.00	0.16	0.00	0.00	0.12	0.00	0.00	0.13	0.00	0.00	0.14	0.00
14	0.00	0.03	0.00	0.00	0.02	0.00	0.00	0.03	0.00	0.00	0.03	0.00